Scalable Startup of Parallel Programs over InfiniBand

WEIKUAN YU, JIESHENG WU AND DHABALESWAR K. PANDA

Technical Report
OSU-CISRC-5/04-TR33

Fast and Scalable Startup of MPI Programs in InfiniBand Clusters*

Weikuan Yu Jiesheng Wu Dhabaleswar K. Panda

Network-Based Computing Lab
Dept. of Computer Science and Engineering
The Ohio State University
{yuw,wuj,panda}@cis.ohio-state.edu

Abstract

Fast and scalable process startup is one of the major challenges in parallel computing over large scale clusters. The startup of a parallel job typically can be divided into two phases: process initiation and connection setup. Both of these phases can become performance bottlenecks. In this paper, we characterize the startup of MPI programs in InfiniBand clusters and identify two startup scalability issues: serialized process initiation in the initiation phase and high communication overhead in the connection setup phase. We propose different approaches to reduce communication overhead and provide fast process initiation. Specifically, to reduce the connection setup time, we have developed one approach with data reassembly to reduce data volume and another with a bootstrap channel to parallelize the communication. Furthermore, we have exploited a process management framework, Multi-purpose Daemons (MPD) system to speed up the process initiation phase. The bootstrap channel is utilized to overcome the scalability limitations of MPD. Our experimental results show that job startup time has been improved by more than 4 times for 128-process jobs in an InfiniBand cluster. Scalability Models derived from these results suggest that the improvement can be more than two orders of magnitudes for the startup of 2048-process jobs.

1. Introduction

The MPI (Message Passing Interface) Standard [4] has evolved as a de facto parallel programming model for distributed memory systems. Traditional research over MPI has been largely focusing on the high performance communication between processes. As cluster computing becomes a prominent platform of high performance computing, scalable process management of MPI applications becomes an active research topic [3, 1]. One of the major challenges in process management is the fast and scalable startup of large-scale applications [2, 7, 10, 5, 9]. This issue becomes even more pronounced in the large scale systems with thousands of nodes. A parallel job is usually launched by a process manager, which is often referred to as the *process initiation phase*. These initiated processes usually require assistance from the process manager to set up peer-to-peer connections before starting communication and computation. This is referred to as the *connection setup phase*.

InfiniBand Architecture (IBA) [8] has been recently standardized by industry to design next generation high-end clusters for both data-center and high performance computing. IBA has been emerging as a popular interconnect for clusters. Large cluster systems with InfiniBand are being deployed. For example, in the Top500 list released in November 2003 [16], the 3rd, 111th, and 116th most powerful supercomputers use InfiniBand as their parallel application communication interconnect. These three systems have 2200, 256, and 512 processors, respectively. The startup of MPI applications in InfiniBand clusters with such a large scale is a challenging issue. It may take more than

*This research is supported in part by a DOE grant #DE-FC02-01ER25506, NSF Grants #EIA-9986052 and #CCR-0204429, and a grant from Los Alamos National Laboratory.

ten minutes to go through the above mentioned process initiation and connection setup phases for an application with 1000 processes without scalable and high performance startup support.

In this paper, we have taken on the challenge to support a scalable and high performance startup of MPI programs over InfiniBand clusters. With MVAPICH [13] as the platform of study, we have analyzed the startup bottlenecks. Accordingly, different approaches are developed to speed up the connection setup phase, one with data reassembly at the process manager and another using pipelined all-to-all broadcast over a ring of InfiniBand QPs (referred to as a bootstrap channel). In addition, we have exploited a process management framework, Multi-purpose Daemons (MPD) system to further speed up the startup. The bootstrap channel is also utilized to reduce the impact of communication bottlenecks in MPD, including multiple process context switches and quadratically increased data volume over the MPD management ring. Over 128 processes, our work improves the startup time by more than 4 times. Scalability Models derived from these results suggest that the improvement can be more than two orders of magnitudes for the startup of 2048-process jobs.

The rest of the paper is structured as follows. Section 2 gives an overview of InfiniBand. Section 3 describes the startup of parallel programs over InfiniBand and the motivation of this study. Section 4 describes the design of startup with different approaches to improve the connection setup time and the process initiation phase. Experiments results are provided in 5. Finally, we conclude the paper in Section 6.

2. Overview of InfiniBand

2.1. Overview of InfiniBand Architecture

The InfiniBand Architecture (IBA) [8] defines a System Area Network (SAN) for interconnecting computing nodes and I/O nodes. It is a communication and management infrastructure supporting both I/O and interprocessor communications (IPC) for one or more computer systems. In an InfiniBand network, a switched communication fabric is defined to allow many devices to communicate concurrently at high bandwidth and low latency. Processing nodes and I/O nodes are connected as end-nodes to the fabric by two kinds of channel adapters: Host Channel Adapters (HCAs) and Target Channel Adapters (TCAs). These IBA channel adapters off-load much of the communication protocol processing from the host CPU, thereby allowing multiple concurrent communications without the traditional overhead. The IBA/SAN provides zero-copy, OS-bypass data movements to its I/O and computing nodes.

The IBA provides a virtual interface to a consumer with the use of Work Queues (WQ). Consumers create work queues in pairs (one for send operations and the other receive operations) and specify its class of service for communication. We call these queue pairs QPs. One or more Completion Queues (CQ) is created and associated with the work queues. The completion of communication requests is reported through completion queues (CQs). The IBA supports both channel and memory semantics. In channel semantics, send/receive operations are used for communication. A receiver must explicitly post a descriptor to receive messages in advance. In memory semantics, RDMA write and RDMA read operations are used. RDMA operations enable the initiator to write data into or read data from memory buffers of the peer side without intervention of the peer side.

2.2. Overview of Reliable Connection Service in InfiniBand

InfiniBand provides four types of transport services: Reliable Connection (RC), Reliable Datagram (RD), Unreliable Connection (UC), and Unreliable Datagram (UD). The often used service is Reliable Connection in the current InfiniBand product and software. It is also our focus of this paper. To support reliable connection service, a connection must be set up between two QPs before any communication.

In the current InfiniBand SDK, each QP has a unique identifier, called *QP-ID*. This is usually an integer. IBA also defines two unicast identifiers [8]: a global identifier (*GID*), and a local identifier (*LID*). GID may be unique across subnets and LID may be unique within a subnet. LID is a 16-bit identifier. To make a connection, both QPs must exchange their QP IDs and LIDs.

3. Motivation

This section provides an overview of MPI program startup in InfiniBand clusters and motivate the study for a scalable startup scheme.

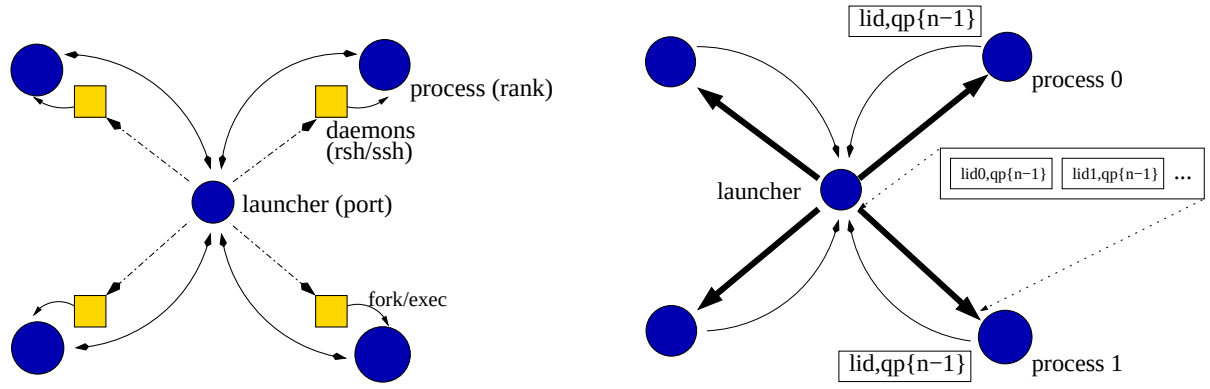
3.1. Startup of MPI Applications using MVAPICH

MVAPICH [12] is a high performance implementation of MPI over InfiniBand. Its design is based on on MPICH [6] and MVICH [11]. The current implementation of MVAPICH utilizes the Reliable Connection (RC) service for the communication between processes.

The connection-oriented feature of IBA RC-based QPs requires each process to create at least one QP for every peer process. To form a fully connected network of N processes, a parallel application needs to create and connect at least $N \times (N-1)$ QPs during the initialization time, Note that it is possible that these QPs can be allocated and connected in a on-demand manner [17], however this requires that the connection management subsystem of IBA can handle asynchronous and client-server model connection establishment, which is not mature yet in the current IBA software. Another reason for the fully-connected connection model is simplicity and robustness. Therefore, this connection model has been used in many MPI implementations, including MVAPICH.

The startup of an MPI application using MVAPICH can also be divided into two phases. As shown in Fig. 1(a), an MPI application using MVAPICH is launched with a simple process launcher iterating over UNIX remote shell (rsh) or secure shell (ssh) to start individual processes. Each process connects back to the launcher via a port exposed by the launcher. Except the process rank, each process has no global knowledge about the parallel program.

In the second phase of connection setup, shown in Fig. 1(b), each process also creates $N-1$ QPs, one for each peer process, given N processes in total. Then, these processes exchange their local identifier (LIDs) and corresponding QP identifiers (QP-IDs), as mentioned in Section 2.2 for connection setup. Since each process is not connected to its peer processes, the exchange of these data has to use the connections that are created to the launcher in the first phase. The launcher collects data about LIDs and QP-IDs from each process, and then sends the combined data back to each process. When having received the combined data about all processes' LIDs and QP-IDs, each process then sets up connections over InfiniBand. A parallel application with fully connected processes is then created.



(a) The Initiation Phase

(b) The Connection Setup Phase

Fig. 1. The Startup of MPI Applications in Current MVAPICH

3.2. The Scalability Problem

The simple startup paradigm described in the earlier subsection is able to handle the startup of any small scale parallel application. However, as the size of an InfiniBand cluster goes to 100s–1000s, the limitation of this paradigm becomes pronounced. For example, launching a parallel application with 2000 processes may take tens of minutes. Among various scalability bottlenecks, there are two main scalability bottlenecks, one in each phase. The first

bottleneck is *rsh/ssh-based startup* in the process initiation phase. This process startup mechanism is simple and straightforward, its performance is very poor on large systems. The second bottleneck is communication overhead for exchanging LIDs and QP-IDs in the connection setup phase. To launch an N-process MPI application, the launcher has to receive data containing $(N-1)$ QP-IDs from each process. Then it returns the combined data with $N \times (N-1)$ QP-IDs to each process. In total, the launcher has to communicate data in the amount of $O(N^3)$ for an N-process application. Each QP-ID is usually a four-byte integer, for a 1024-process application the launcher will receive almost 4 MegaBytes data and sends almost 4 Gigabytes of data. This communication usually goes through the management network which is normally Fast Ethernet or Gigabit Ethernet. Significant communication overhead incurs to slow down the application startup.

In this paper, we plan to address these two startup scalability problems in InfiniBand clusters. We focus on different approaches to reduce communication overhead in the connection setup phase and fast process initiation in the initiation phase.

4. Designing Scalable Startup Schemes

This section describes the design of scalable startup schemes in InfiniBand clusters. We first describe different approaches used to enhance the connection setup phase while the processes are still launched via rsh/ssh daemons. Then we exploit the advantages of a process management framework, *MPD* [3], to replace the rsh/ssh based scheme and achieve efficient process initiation. We also characterize some MPD features and their limitations to the requirement of MPI application startup in InfiniBand clusters. A bootstrap channel is introduced to overcome these limitations.

4.1. Efficient Connection Setup

As mentioned in the previous section, because the launcher has to collect, combine and broadcast QP IDs, the volume of these data scales up in the order of $O(N^3)$, which leads to prolonged connection setup time. One needs to consider two directions in order to reduce the connection setup time. The first direction is to reduce the volume of data that needs to be communicated. The other direction is to parallelize communication for the exchange of QP ID's. Along these two directions, we propose the following two approaches to improve the connection setup phase, respectively.

4.1.1. Approach 1: Reducing the Data Volume with Data Reassembly (DR)

To have processes fully connected over InfiniBand, each process needs to connect with another peer process via one QP. This means that, each process needs to know $N-1$ QP IDs, one from each peer process, So out of the combined data about $N \times (N-1)$ QP IDs in the launcher, one process only needs to receive N QPs that is specifically targeted for it. This requires a centralized component, i.e., the launcher, to collect and reassembly QP IDs. The biggest advantage of this data reassembly (*DR*) scheme is that the data volume can be reduced to the order of $O(N^2)$. But there are several disadvantages associated with this scheme. First, the entire data of QP IDs need to be reassembled at the launcher, possibly multiple times. This leads to another performance/scalability bottleneck. Second, the total exchange of QP IDs is sequentialized at the launcher. The launcher must receive all QP IDs first and then reassemble QP IDs and send the reassembled QP IDs to each process. The receive-reassembly-send procedure is serialized.

4.1.2. Approach 2: Parallelizing Communication with a Bootstrap Channel (BC)

In order to parallelize the communication, one must find out the available parallelism and the algorithm to exploit that. Reexamination of the startup can shed insights on these issues. Essentially, what needs to be achieved at the startup time is an all-to-all personalized exchange of QP IDs, i.e., each process receives the specific QP IDs from other processes. In the original startup scheme as shown in Figure 1, the launcher performs a gather/broadcast to help the all-to-all broadcast of their QP data. On top of that, the DR scheme in Section 4.1.1 reassembles and "personalizes" QP data to reduce the data volume. Both do not exploit the parallelism of all-to-all personalized exchange. Algorithms that parallelize an all-to-all personalized exchange can be used here. These algorithms are usually based on a ring-, hypercube- or torus-based topology, which requires more connections to be provided among processes. With the

initial star topology in the original startup scheme, providing these connections has to be done through the launcher. However, since a parallelization algorithm can potentially overlap both sending and receiving QP data, it promises better scalability over clusters with larger sizes.

Among the three possible parallel topologies, the ring-based topology requires the least number of additional connections, i.e., 2 per process. This would minimize the impact of the ring setup time. Another design option to be considered is that which type of connections should be provided. Either TCP/IP- or InfiniBand-based connections can be used. Since the communication over InfiniBand is much faster than that over TCP/IP, as shown in Fig. 2. a ring of InfiniBand QPs is the ideal choice to exploit parallelism, and speed up communication.

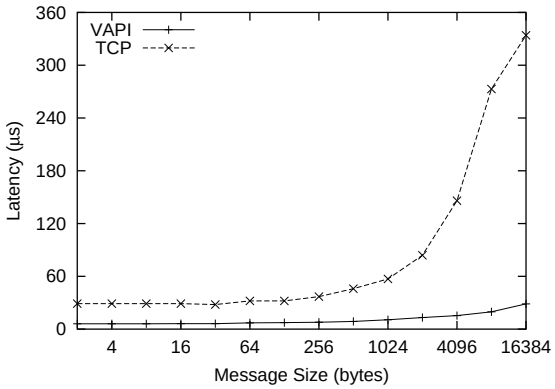
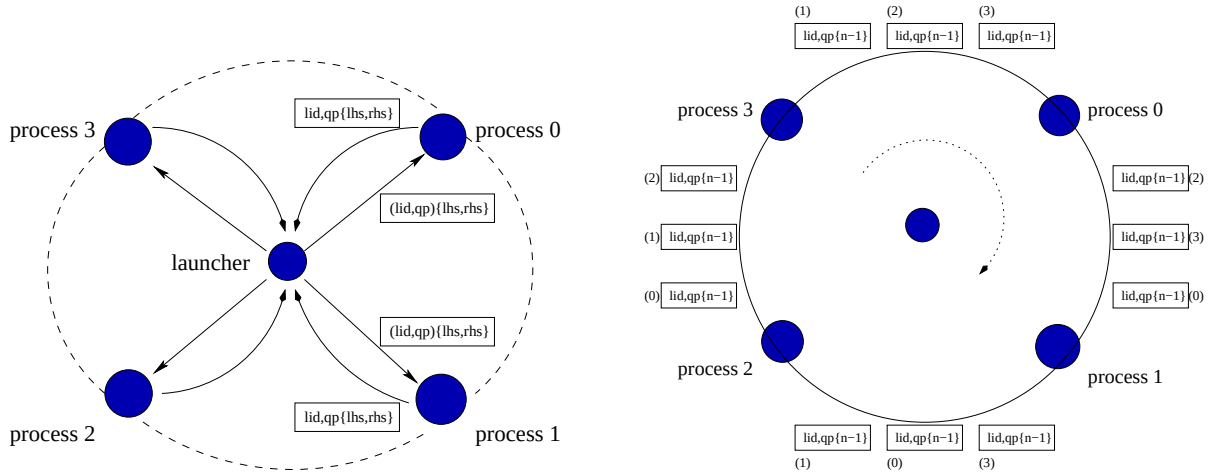


Fig. 2. Latency Comparison of Communication over TCP/IP on Fast Ethernet and VAPI on InfiniBand

In our design, we choose a ring of InfiniBand QPs. The second approach works as follows. First, each process creates two QPs for its left hand side (lhs) and right hand side (rhs) processes, respectively. We call these QPs *bootstrap QPs*. Second, the DR scheme mentioned in Section 4.1.1 is used to set up connections between these bootstrap QPs as shown in Figure 3(a). Thus, a ring of connections over InfiniBand is created, as shown by the dotted line in Figure 3(a). We refer to this ring as a *bootstrap channel*. After the bootstrap channel is set up, each process initiates a broadcast of its own QP IDs over the bootstrap channel in the clockwise direction as shown in Fig. 3(b) with four processes. Each process also forwards what it receives to its next process. In this scheme, we take advantage of both communication parallelism and high performance of bootstrap channel over InfiniBand to reduce the communication overhead for the connection setup.



(a) Setup a Bootstrap Channel

(b) Ring-Based All-to-all Broadcast of QP Data

Fig. 3. Parallelizing the Total Exchange of InfiniBand Queue Pair Data

4.2. Fast Process Initiation with MPD

The rsh/ssh-based process initiation is another bottleneck of the application startup in large clusters. It is desirable that we take advantage of some resource management systems [14, 2, 9, 15, 3] to facilitate the process initiation. MPD [3] is designed to be a general process manager interface that provide the needed support for MPICH, from which MVAPICH is developed. It mainly provides fast startup of parallel applications and convenient runtime management of parallel jobs. Instead of sequentially launching processes over rsh/ssh daemons, MPD parallelizes the launching of application processes over a ring of MPD daemons. As shown in Fig. 4, a console process from the user front end sends a request to the ring of MPD daemons to launch a parallel job. This request is instantly propagated across the ring. The specified set of daemons then start a ring of management processes, each of which in turn starts an application process for the parallel job. MPD achieves the fast startup of parallel jobs because the startup time for all pairs of managers and application processes is parallelized. An explicit interface (i.e. BNR) is also introduced to help the parallel programming library to obtain information from the process manager (in this case, MPD). This type of information is typically only stored and known in the MPD database, such as the rank of an individual process. In addition, processes can exchange information through its put/fence/get model, exposed as the BNR interface. One process stores (puts) a (key,value) pair at its manager, a part of the MPD database, then another process retrieves (gets) that value by providing the same key.

Although this fast, parallelized process startup from MPD solves the process initiation problem, the significant volume of QP data poses a great challenge to the MPD model. As shown in Fig. 5(a), the database is distributed over the ring of manager processes when each process stores (puts) their process-specific data to its manager. To collect the data from every peer process, one process has to send a request and get the reply back for the target process. At the completion of these data exchanges, each process then sets up connections with all the peers, as the process 0 shown in Fig. 5(a). Together, messages for the request and the reply make a complete round over the manager ring. For a parallel job with N processes, there are $N \times (N-1)$ message exchanges in total. Each of these messages is in the order of $O(N)$ bytes and has to go through the ring of manager processes. In addition, since application processes store and retrieve data through their corresponding manager processes at each node, process context switches are very frequent and they further degrade the performance of ring-based communication. Furthermore, the message passing is over TCP/IP sockets, which delivers lower performance than InfiniBand-based connections as shown in Fig 2.

There are different alternatives to overcome these limitations. One way of doing that is to replace the connections for the MPD manager ring with VAPI connections to provide fast communications. In addition, copies of QP data can be saved at each manager process as the first copy of QP data passes through the ring. Then further retrieve

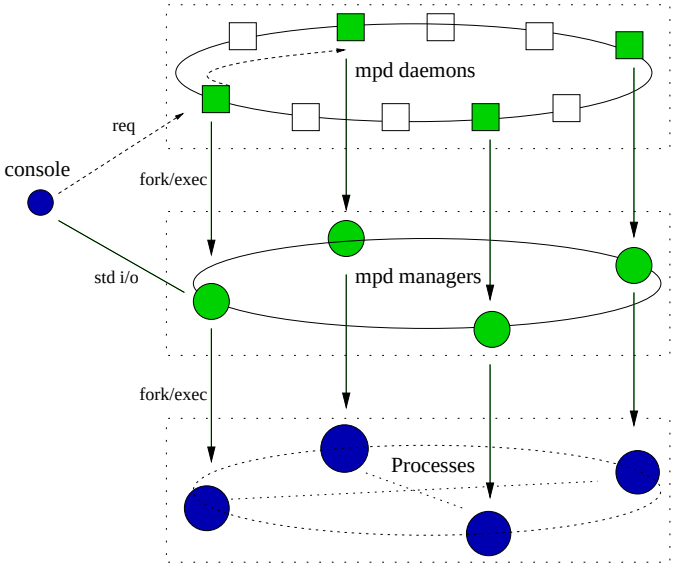
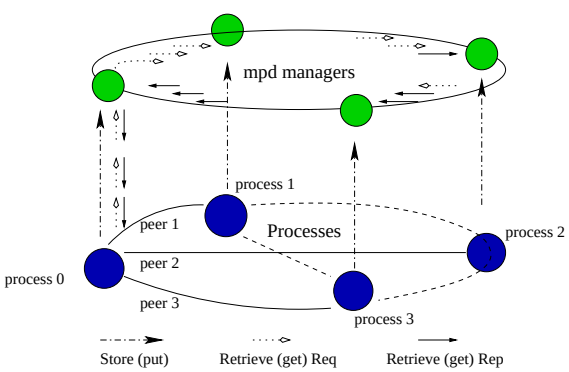
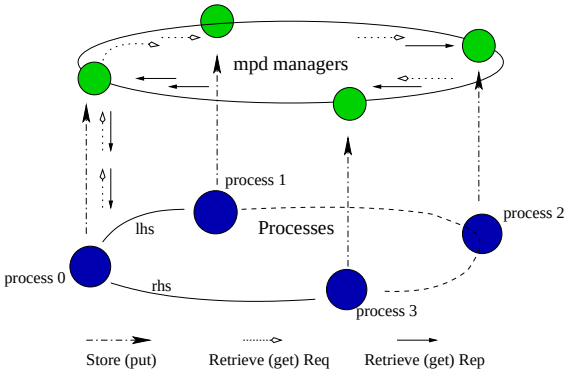


Fig. 4. MPD Parallelized Process Initialization



(a) Exchange of Queue Pair IDs Over the Ring of Manager



(b) Setting up Bootstrap Channel within Processes

Fig. 5. Improving the Scalability of MPD-Based Startup

(get) requests can get the data from the local manager directly instead of the MPD manager ring. This approach will improve the communication time, however, the process context switches still exist between the application processes and manager processes. In addition, retrieve requests made before QP data reaches the local manager process still has to go through the manager ring. Last but not least, this approach necessitates a significant amount of instrumentation of MPD code and can affect its portability.

Instead of exchanging all the QP data over the ring of MPD manager processes, we propose to perform the exchange of QP IDs over the bootstrap channel described in Section 4.1.2. However, setting up the bootstrap channel still has to go through the ring of manager processes. As shown in Fig. 5(b), each process first creates and stores QP IDs for its left side (lhs) and right hand side (rhs) processes to the local manager. Then, from the database, they retrieve the data for its left hand side and right hand side processes, and use these data to set up InfiniBand connections. Eventually a ring of such connections are constructed and form a bootstrap channel. This bootstrap channel is then utilized to perform a complete exchange of QP IDs as described in Section 4.1.2. Since this bootstrap channel is provided within the application processes and over InfiniBand, this approach will not only provide fast communication performance and eliminate the process context switches, but also reduce the number of communications on each manager process.

5. Performance Evaluation

In this section we describe experimental results of our implementation. Our implementation is based on MVAPICH 0.9.1. We provide both the rsh/ssh-based and MPD-based parallel startup. For the rsh/ssh-based startup, we provide fast connection setup with both the data reassembly (DR) and the bootstrap channel (BC)-based approaches. For MPD-based startup, we use the bootstrap channel inside the application processes to enhance the startup time.

Experiments were conducted on a 256-node cluster of 4GB DRAM dual-SMP 2.4GHz Xeon at the Ohio Supercomputing Center. For fast network discovery with data reassembly (DR) or the bootstrap channel (BC), we used ssh to launch the parallel processes. Performance comparisons were provided against MVAPICH 0.9.1 (Original). To eliminate the impact of network communication to file access, (file system performance could be a big bottleneck for a large cluster), all binary executable files were duplicated at local disks.

Number of Processes	4	8	16	32	64	128
Original (sec)	0.59	0.92	1.74	3.41	7.3	13.7
SSH-DR (sec)	0.58	0.94	1.69	3.37	6.77	13.45
SSH-BC (sec)	0.61	0.95	1.70	3.38	6.76	13.3
MPD-BC (sec)	0.61	0.63	0.64	0.84	1.58	3.10

Table 1. Comparisons of Parallel Job Startup Time over MVAPICH with Different Approaches

5.1. Experimental Results

Table 1 shows the startup time for parallel jobs of different number processes using different approaches. SSH-DR represents ssh-based startup with QP data assembly (DR) at the process launcher. SSH-BC represents ssh-based startup using the bootstrap channel (BC) to exchange QP IDs. MPD-BC represents MPD-based startup with a bootstrap channel for the QP Id exchange.

As the number of processes increases, both SSH-DR and SSH-BC reduce the startup time, compared to the original approach, This is because data reassembly can reduce the data volume by an order of $O(N)$ and the bootstrap channel can parallelize the communication time. Note that BC-based approach performs slightly worse than the the original and DR-based approach for small number of processes. This is due to the overhead from setting up the additional ring over InfiniBand. As the number of processes increases, the benefits becomes greater. However, with less than or equal to 128 processes, the length of entire QP IDs is still less than 2K. With either TCP-based or VAPI-based communication, the communication time for the connection setup, is in the order of milliseconds, while the total startup time is in the order of seconds, dominated by the process initiation time. However, either SSH-BC and SSH-DR will be able to provide more scalable startup for a job with thousands of processes since they both are able to remove the major communication bottleneck imposed by the sheer volume QP data. In contrast, the MPD-based approach with a bootstrap channel provides the most scalable startup. On one hand, MPD-BC provides efficient parallelized process initialization, compared to the ssh-based schemes. On the other hand, it also pipelines the QP data exchange over a ring of VAPI connections, hence this approach speeds up the connection setup phase. Compared to the original approach, the MPD-BC approach reduces the startup time for a 128-process job by more than 4 times.

5.2. Analytical Models and Evaluations for Large Clusters

As indicated by the results from Section 5, the benefits of the designed schemes will be more pronounced for parallel jobs with larger number of processes. In this section, we further analyze the performance of different startup schemes and provide parametrized models to gain insights about their scalability over large clusters. The total startup time $T_{startup}$ can be divided into two components: the process initiation time and the connection setup time, denoted as T_{init} and T_{conn} respectively. Based on the scalability analysis, we use the following model to describe the startup time of the following schemes, the original scheme (Original), ssh-based scheme with data reassembly (SSH-DR) and the MPD-based scheme with the bootstrap channel (MPD-BC). Each of the models shows the time for the startup of

N processes, and the last component describes the time for other overheads that are not quantified in the models, for example, process switching overhead.

Original: $T_{startup} = (O_0 * N) + (O_1 * N * (W_N + W_{N^2})) + O_2$

The process initiation phase time T_{init} scales linearly as the number of processes increases with ssh/rsh-based approaches, while during the connection setup there are 2N messages communicated over TCP/IP. Half of them are gathered by the launcher, each being in the order of $O(N)$ bytes; the other half are scattered by the launcher, each of $O(N^2)$ bytes .

SSH-DR: $T_{startup} = (D_0 * N) + (D_{comp} * N^3 + D_1 * 2N * W_N) + D_2$

The process initiation time T_{init} scales linearly with ssh/rsh. During the connection setup phase, the amount of computation scales in the order of $O(N^3)$ (the constant D_{comp} can be very small, being the time for extracting one QP Id), and there are 2*N message communicated over TCP/IP. Half of them are gathered by the launcher, each being in the order of $O(N)$ bytes; The other half are scattered by the launcher, each of them is only $O(N)$ bytes due to reassembly.

MPD-BC: $T_{startup} = (M_0 + N * W_{req}) + (M_{ch\_setup} * N + M_1 * N * W_N) + M_2$

The process initiation time T_{init} scales constantly using MPD, however there is a small fractional increase of communication time for the request message W_{req}. During the connection setup phase, the time to setup a bootstrap channel increases in the order of $O(N)$. and also each process handles N message in the pipeline, each in the order of $O(N)$ bytes.

Original: $T_{startup}$ (sec) $= (0.100 * N) + (10.5 * N * (W_N + W_{N^2})) + 0.12$

SSH-DR: $T_{startup}$ (sec) $= (0.100 * N) + (8.5e^{-9} * N^3 + 10.5 * N * W_N) + 0.12$

MPD-BC: $T_{startup}$ (sec) $= (0.20 + 0.0010 * N) + (0.0180 * N + 2.5 * N * W_N) + 0.30$

The above scalability models are parameterized based on our analytical modeling. As shown in Fig. 6, the experiment results confirm the validity of these models for jobs with 4 to 128 processes. Fig. 7 shows the scalability of different startup schemes when applying the same models to larger jobs from 4 to 2048 processes. Both SSH-DR and MPD-BC improves the scalability of job startup significantly. It is to note that, MPD-BC scheme improves the startup time by about two orders of magnitudes for 2048-process jobs.

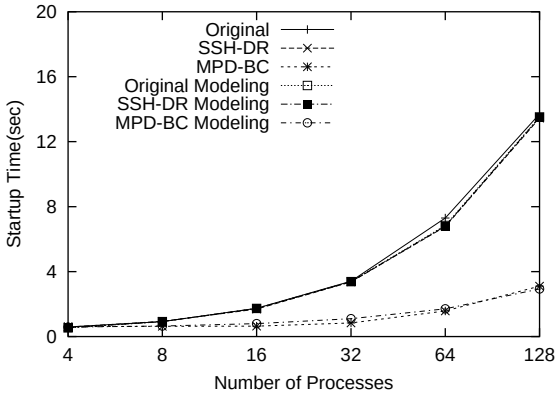


Fig. 6. Performance Modeling of Different Startup Schemes

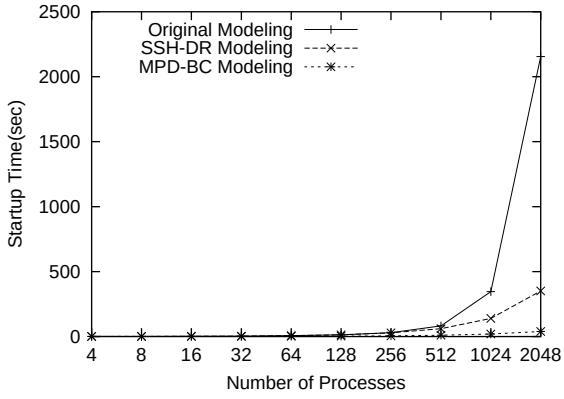


Fig. 7. Scalability Comparisons of Different Startup Schemes

6. Conclusions and Future Work

In this paper, we have presented schemes to support scalable startup of MPI programs in InfiniBand clusters. With MVAPICH as the platform of study, we have characterized the startup of MPI jobs into two phases: process initiation

and connection setup. To speed up connection setup phase, we have developed two approaches, one with queue pair data reassembly at the launcher and the other with a bootstrap channel. In addition, we have exploited a process management framework, Multi-purpose Daemons (MPD) system, to improve the process initiation phase. The performance limitations in the MPD's ring-based data exchange model, such as exponentially increased communication time and numerous process context switches, are eliminated by using the proposed bootstrap channel. We have implemented these schemes in MVAPICH [13]. Our experimental results show that, for 128-process jobs, the startup time has been reduced by more than 4 times. We have also developed an analytical model to project the scalability of the startup schemes. The derived models suggest that the improvement can be more than two orders of magnitudes for the startup of 2048-process jobs with the MPD-BC startup scheme.

In future, we want to provide a file broadcast mechanism to MPD system to achieve efficient loading of jobs [5, 10]. We also plan to expand the BNR interface of MPD system so that the manager database can be suitably distributed according to the hints provided in the requests. Furthermore, we intend to provide a hypercube-based scalable startup over really large systems, e.g., future Peta-scale clusters with tens of thousands of processors.

Software Availability

The new startup code will be released in the next release of MVAPICH version 0.9.5, and be available at the link: http://nowlab.cis.ohio-state.edu/projects/mpi-iba/.

References

[1] M. Baker, G. Fox, and H. Yau. Cluster Computing Review, November 1995.

[2] R. Brightwell and L. A. Fisk. Scalable parallel application launch on Cplant. In *Proceedings of Supercomputing, 2001*, Denver, Colorado, November 2001.

[3] R. Butler, W. Gropp, and E. Lusk. Components and interfaces of a process management system for parallel programs. *Parallel Computing*, 27(11):1417–1429, 2001.

[4] M. P. I. Forum. MPI: A message-passing interface standard. *The International Journal of Supercomputer Applications and High Performance Computing*, 8(3–4), 1994.

[5] E. Frachtenberg, F. Petrini, J. Fernandez, S. Pakin, and S. Coll. STORM: Lightning-Fast Resource Management. In *Proceedings of the Supercomputing '02*, Baltimore, MD, November 2002.

[6] W. Gropp, E. Lusk, N. Doss, and A. Skjellum. A High-Performance, Portable Implementation of the MPI Message Passing Interface Standard. *Parallel Computing*, 22(6):789–828, 1996.

[7] E. Hendriks. Bproc: The beowulf distributed process space. In *Proceedings of the 16th Annual ACM International Conference on Supercomputing (ICS 02)*, New York, New York, June 2002.

[8] Infiniband Trade Association. http://www.infinibandta.org.

[9] M. Jette and M. Grondona. SLURM: Simple Linux Utility for Resource Management. In *Proceedings of the 4th LCI International Conference on Linux Clusters*, San Jose, CA, June 2003.

[10] A. Kavas, D. Er-El, and D. G. Feitelson. Using Multicast to Pre-Load Jobs on the ParPar Cluster. *Parallel Computing*, 27(3):315–327, 2001.

[11] Lawrence Berkeley National Laboratory. MVICH: MPI for Virtual Interface Architecture. `http://www.nersc.gov/research/FTG/mvich/ index.html`, August 2001.

[12] J. Liu, J. Wu, S. P. Kini, D. Buntinas, W. Yu, B. Chandrasekaran, R. Noronha, P. Wyckoff, and D. K. Panda. MPI over InfiniBand: Early Experiences. Technical Report OSU-CISRC-07/00-TR16, Department of Computer and Information Sci., The Ohio State University, Columbus, OH 43210, January 2003.

[13] Network-Based Computing Laboratory. MVAPICH: MPI for InfiniBand on VAPI Layer. `http://nowlab.cis.ohio-state.edu/ projects/mpi-iba/index.h tml`, January 2003.

[14] OpenPBS Documentation. http://www.openpbs.org/docs.html.

[15] Quadrics Supercomputers World, Ltd. Quadrics Documentation Collection. http://www.quadrics.com/onlinedocs/Linux/html/index.html.

[16] TOP 500 Supercomputers. http://www.top500.org/, 2003.

[17] J. Wu, J. Liu, P. Wyckoff, and D. K. Panda. Impact of On-Demand Connection Management in MPI over VIA. In *Proceedings of the IEEE International Conference on Cluster Com puting*, 2002.